Discussion of Latouche and Lee

Guido Consonni

Università Cattolica del Sacro Cuore, Milan

Pierre Latouche
Bayesian Variable Selection for Globally Sparse Probabilistic PCA

Unsupervised feature selection
Quoting from Authors' paper

Pierre Latouche
Bayesian Variable Selection for Globally Sparse Probabilistic PCA

Unsupervised feature selection
Quoting from Authors' paper

- hazy and exciting problem

Pierre Latouche
Bayesian Variable Selection for Globally Sparse Probabilistic PCA

Unsupervised feature selection
Quoting from Authors' paper

- hazy and exciting problem
- ill-posed when no specific learning task (such as clustering) is driving it

Sparse PCA

- Sparse PCA (Zhou, Hastie, Tibshirani, 2006)
  Principal components are sparse (contain only few variables)
- Principal components do *not* have the same sparsity pattern
  (same active variables)
  Each component has to be interpreted individually
  - good for visualization
  - good for interpretation (with some luck)

Feature/variable selection

- Sparse PCA is bad for variable selection
  (each principal component has its own sparsity pattern)
- Leading idea of the paper
  project data onto a
  globally sparse subspace
  (space spanned by vectors with the same sparsity pattern)

Globally Sparse Probabilistic PCA

$$x_i = VWy_i + \bar{V}\epsilon_{1i} + V\epsilon_{2i}; \quad i = 1, \ldots, n$$

- prior on W

$$w_{ij} \mid \alpha \overset{iid}{\sim} N(0, 1/\alpha^2)$$

- For $\sigma_2 \to 0$ closed form expression for the integrated likelihood

$$p(x_i \mid v, \alpha, \sigma_1^2)$$

- Estimate
  u (a continuous relaxation of v)
  α
  σ_1^2

  using variational EM (VEM)

Questions

- Main idea of the paper
  project data onto a globally sparse subspace
  while
  preserving a large part of the variance

Questions

- Main idea of the paper
  project data onto a globally sparse subspace
  while
  preserving a large part of the variance
  Former is clearly modeled
  How is the latter incorporated transparently into the model?

Questions

- Main idea of the paper
  project data onto a globally sparse subspace
  while
  preserving a large part of the variance
  Former is clearly modeled
  How is the latter incorporated transparently into the model?
- Current fully Bayes variable selection treats the discrete indicator $v \in \{0,1\}^p$ as random with a product Bernoulli- Beta prior
  Advantages of posterior distribution on v
  - model uncertainty
  - posterior probability of variable inclusion
  - multiplicity correction
  - model averaging

Questions

- Main idea of the paper
  project data onto a globally sparse subspace
  while
  preserving a large part of the variance
  Former is clearly modeled
  How is the latter incorporated transparently into the model?
- Current fully Bayes variable selection treats the discrete indicator $v \in \{0,1\}^p$ as random with a product Bernoulli- Beta prior
  Advantages of posterior distribution on v
  - model uncertainty
  - posterior probability of variable inclusion
  - multiplicity correction
  - model averaging
- You opt for *estimating* v (EB style)

Questions

- Main idea of the paper
  project data onto a globally sparse subspace
  while
  preserving a large part of the variance
  Former is clearly modeled
  How is the latter incorporated transparently into the model?
- Current fully Bayes variable selection treats the discrete indicator $v \in \{0,1\}^p$ as random with a product Bernoulli- Beta prior
  Advantages of posterior distribution on v
  - model uncertainty
  - posterior probability of variable inclusion
  - multiplicity correction
  - model averaging
- You opt for *estimating* v (EB style)
  The computational advantage is clear

Questions

- Main idea of the paper
  project data onto a globally sparse subspace
  while
  preserving a large part of the variance
  Former is clearly modeled
  How is the latter incorporated transparently into the model?
- Current fully Bayes variable selection treats the discrete indicator $v \in \{0,1\}^p$ as random with a product Bernoulli- Beta prior
  Advantages of posterior distribution on v
  - model uncertainty
  - posterior probability of variable inclusion
  - multiplicity correction
  - model averaging
- You opt for *estimating* v (EB style)
  The computational advantage is clear
  However the price to be paid seems important
  Model uncertainty is no longer available

An Aside: Graphical Models

$$x_i \mid \Sigma \overset{iid}{\sim} N_p(0, \Sigma)$$

Assume that the density of x_i factorizes according to a directed acyclic graph $\mathcal{D}$

Distribution is Markov wrt $\mathcal{D}$

An Aside: Graphical Models

$$x_i \mid \Sigma \overset{iid}{\sim} N_p(0, \Sigma)$$

Assume that the density of x_i factorizes according to a directed acyclic graph $\mathcal{D}$

Distribution is Markov wrt $\mathcal{D}$

Two possible goals

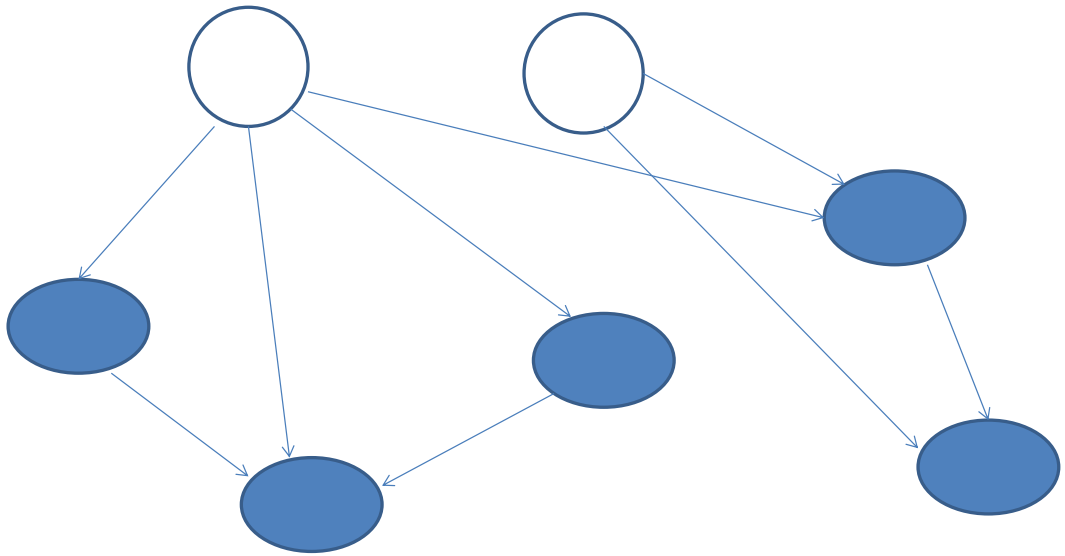
- structural learning of $\mathcal{D}$
  (or rather its Markov equivalence class)
- estimating causal effect on Y of an intervention on variable X_j
  accounting for model uncertainty

Structural Learning of Graphs with Latent Variables

- With latent variables learning $\mathcal{D}$ becomes very hard
- Progress can be made if only a few latent variables are assumed to have a direct effect on observed variables and the latent variables have no parents
- Model

$$x_i = VWy_i + \epsilon_i$$

$y_i \sim N_d(0, I_d)$; $\epsilon_i \sim N(0, \Sigma)$
Σ Markov wrt $\mathcal{D}$



- $p \geq n$ setting
- popular to reduce number of effective parameters
- especially if there is a natural ordering among variables

Estimating a banded covariance

- Current frequentist methods
  do not provide interval estimators for functionals of covariance matrices
  do not achieve optimal, or nearly optimal, minimax rate
- Bayesian methods
  Scarce
  Difficult to place a prior because of complex parameter space
  Covariance graph model by Khare *et al* (2011) not suitable in high-dimensions

Post-processed posteriors

- Conceptually straightforward
- Computationally fast
- *Initial prior*
- *Initial posterior*
- *Post-processed posterior*

Post-processed posteriors

- Conceptually straightforward
- Computationally fast
- *Initial prior*
- *Initial posterior*
- *Post-processed posterior*

Corresponds (almost) to a *data-dependent* prior π^{pp}

$$\pi^{pp}\{\Sigma(\theta_1, 0)\} \propto \int \pi^{i}\{\Sigma(\theta_1, \theta_2)\} \frac{p(X_n \mid \Sigma(\theta_1, \theta_2))}{p(X_n \mid \Sigma(\theta_1, 0))} d\theta_2$$

Post-processed posteriors

- Conceptually straightforward
- Computationally fast
- *Initial prior*
- *Initial posterior*
- *Post-processed posterior*

Corresponds (almost) to a *data-dependent* prior π^{pp}

$$\pi^{pp}\{\Sigma(\theta_1, 0)\} \propto \int \pi^{i}\{\Sigma(\theta_1, \theta_2)\} \frac{p(X_n \mid \Sigma(\theta_1, \theta_2))}{p(X_n \mid \Sigma(\theta_1, 0))} d\theta_2$$

Puts the method on firm Bayesian ground

Questions

- Method is reasonable
  if the prior is reasonable

Questions

- Method is reasonable
  if the prior is reasonable
  What's the intuition behind the structure of π^{pp}?
- Difference between marginals of diagonal entries under the initial prior (Inverse Wishart) and corresponding marginals under the post-processed prior are minimal in the example for $p = 2$
  Larger p?
- Robustness of π^{pp}

Decision-theoretic setting

- Action: Post-processed posterior

Decision-theoretic setting

- Action: Post-processed posterior
- Statistical procedure
  Pair=(Initial prior, post-processing function)

Decision-theoretic setting

- Action: Post-processed posterior
- Statistical procedure
  Pair=(Initial prior, post-processing function)
- P-loss
- P-risk

Convergence rates

- Banded post-processed posterior is nearly optimal
  It has, up to a $(\log k)^2$ factor, the minimax rate
- This result also applies to "ordinary" priors on banded covariances
  (post-processing function=identity)

Beyond the Wishart

- The initial prior of the paper is the Inverse Wishart (IW)
- What's special about the IW?
- Conjugacy?

DAG-Wishart prior

- Generalizes ordinary Wishart prior
  - precision matrices Markov wrt a DAG
    If DAG is complete it is also a prior on standard spd matrices
  - Allows multiple shape hyper-parameters
    Flexible/enriched Wishart

DAG-Wishart prior

- Generalizes ordinary Wishart prior
  - precision matrices Markov wrt a DAG
    If DAG is complete it is also a prior on standard spd matrices
  - Allows multiple shape hyper-parameters
    Flexible/enriched Wishart
- Σ: covariance matrix; $\Omega = \Sigma^{-1}$
  $\Omega = LD^{-1}L\top$: modified Cholesky decomposition
  $\mathcal{D}$: DAG
  Ω Markov wrt $\mathcal{D} \Leftrightarrow L_{ij} = 0$ whenever $i \notin pa_j(\mathcal{D})$

DAG-Wishart prior

- Generalizes ordinary Wishart prior
  - precision matrices Markov wrt a DAG
    If DAG is complete it is also a prior on standard spd matrices
  - Allows multiple shape hyper-parameters
    Flexible/enriched Wishart
- Σ: covariance matrix; $\Omega = \Sigma^{-1}$
  $\Omega = LD^{-1}L\top$: modified Cholesky decomposition
  $\mathcal{D}$: DAG
  Ω Markov wrt $\mathcal{D} \Leftrightarrow L_{ij} = 0$ whenever $i \notin pa_j(\mathcal{D})$

$$\pi_{\mathcal{D}}(D, L \mid U, \{\alpha_i\}) \propto \exp\left\{-\frac{1}{2} tr(LD^{-1}L\top)U\right\} \prod_i D_{ii}^{-\frac{\alpha_i}{2}}$$

DAG-Wishart prior

- Generalizes ordinary Wishart prior
  - precision matrices Markov wrt a DAG
    If DAG is complete it is also a prior on standard spd matrices
  - Allows multiple shape hyper-parameters
    Flexible/enriched Wishart
- Σ: covariance matrix; $\Omega = \Sigma^{-1}$
  $\Omega = LD^{-1}L\top$: modified Cholesky decomposition
  $\mathcal{D}$: DAG
  Ω Markov wrt $\mathcal{D} \Leftrightarrow L_{ij} = 0$ whenever $i \notin pa_j(\mathcal{D})$

$$\pi_{\mathcal{D}}(D, L \mid U, \{\alpha_i\}) \propto \exp\left\{-\frac{1}{2} tr(LD^{-1}L\top)U\right\} \prod_i D_{ii}^{-\frac{\alpha_i}{2}}$$

- Preserves conjugacy

DAG-Wishart prior

- Generalizes ordinary Wishart prior
  - precision matrices Markov wrt a DAG
    If DAG is complete it is also a prior on standard spd matrices
  - Allows multiple shape hyper-parameters
    Flexible/enriched Wishart
- Σ: covariance matrix; $\Omega = \Sigma^{-1}$
  $\Omega = LD^{-1}L\top$: modified Cholesky decomposition
  $\mathcal{D}$: DAG
  Ω Markov wrt $\mathcal{D} \Leftrightarrow L_{ij} = 0$ whenever $i \notin pa_j(\mathcal{D})$

$$\pi_{\mathcal{D}}(D, L \mid U, \{\alpha_i\}) \propto \exp\left\{-\frac{1}{2} tr(LD^{-1}L\top)U\right\} \prod_i D_{ii}^{-\frac{\alpha_i}{2}}$$

- Preserves conjugacy
- Conjecture
  DAG-Wishart on Ω under a *complete* DAG as initial prior
  might achieve better finite-sample performance in terms of P-risk than IW.

THANKS TO PIERRE AND JAEYONG !